Understanding Trust and Reliance Development in AI Advice: Assessing Model Accuracy, Model Explanations, and Experiences from Previous Interactions.

PATRICIA K. KAHR, Eindhoven University of Technology, The Netherlands
GERRIT ROOKS, Eindhoven University of Technology, The Netherlands
MARTIJN C. WILLEMSEN, Eindhoven University of Technology, The Netherlands and Jheronimus Academy of Data Science, The Netherlands
CHRIS C. P. SNIJDERS, Eindhoven University of Technology, The Netherlands

People are increasingly interacting with AI systems, but successful interactions depend on people trusting these systems only when appropriate. Since neither gaining trust in AI advice nor restoring lost trust after AI mistakes is warranted, we seek to better understand the development of trust and reliance in sequential human-AI interaction scenarios. In a 2x2 between-subject simulated AI experiment, we tested how model accuracy (high vs. low) and explanation type (human-like vs. abstract) affect trust and reliance on AI advice for repeated interactions. In the experiment, participants estimated jail times for 20 criminal law cases, first without and then with AI advice. Our results show that trust and reliance are significantly higher for high model accuracy. In addition, reliance does not decline over the trial sequence, and trust increases significantly with high accuracy. Human-like (vs. abstract) explanations only increased reliance on the high-accuracy condition. We furthermore tested the extent to which trust and reliance in a trial round can be explained by trust and reliance experiences from prior rounds. We find that trust assessments in prior trials correlate with trust in subsequent ones. We also find that the cumulative trust experience of a person in all earlier trial rounds correlates with trust in subsequent ones. Furthermore, we find that the two trust measures, trust and reliance, impact each other: prior trust beliefs not only influence subsequent trust beliefs but likewise influence subsequent reliance behavior, and vice versa. Executing a replication study yielded comparable results to our original study, thereby enhancing the validity of our findings.

CCS Concepts: • **Human-centered computing → User studies**; **HCI theory, concepts and models**.

Additional Key Words and Phrases: Trust & Reliance in AI Over Time, Willingness to Follow AI Advice, Collaborative Decision-Making

1 INTRODUCTION

It is likely that humans will interact increasingly with artificially intelligent (AI) systems. What may appear to be an exciting opportunity is tempered by the reality that significant issues still exist: human-algorithm interaction (HAI) is not always designed appropriately, and, hence, AI systems are not yet widely accepted. In principle, people need to trust that an AI system can improve their decision-making, but often, they do not trust the support an AI

Authors' addresses: Patricia K. Kahr, p.k.kahr@tue.nl, Eindhoven University of Technology, Groene Loper 3, P.O. Box 513, Eindhoven, The Netherlands, 5612 AE; Gerrit Rooks, Eindhoven University of Technology, Eindhoven, The Netherlands, g.rooks@tue.nl; Martijn C. Willemsen, Eindhoven University of Technology, Groene Loper 3, P.O. Box 513, Eindhoven, The Netherlands, 5612 AE and Jheronimus Academy of Data Science, 5211 DA 's-Hertogenbosch, The Netherlands, 5211 DA, m.c.willemsen@tue.nl; Chris C. P. Snijders, Eindhoven University of Technology, Eindhoven, The Netherlands, c.c.p.snijders@tue.nl.

ACM 2160-6463/2024/8-ART
https://doi.org/10.1145/3686164

system offers [61]. Trust can be impeded by a user's attitudes or traits, the AI system's properties or its modalities of interacting with the user, and the context of a decision. Trust may require specific knowledge or training, and people may have to familiarize themselves with AI to learn when they can or cannot rely on its results. Also, they should trust AI to the right extent and not over-trust a failing system or under-trust an optimal one. The current research landscape on trust in AI covers several aspects: the influence of system properties [30, 85, 92], user traits [6, 18, 61], the interaction modalities between the user and an AI system [9, 21, 28], or the task context of human-AI-decision-making [88]. Despite the volume and variety of existing research, findings regarding trust in HAI are often inconclusive. Much of the trust in AI research is based on laboratory experiments with short-term experimental setups that allow limited ecological validity and make it difficult to draw conclusions for real-world applications with higher complexity [37]. In addition, studies tend to investigate HAI with single-shot studies rather than considering HAI in repeated rounds of interaction, cf. [37, 97]. Furthermore, work that measures repeated interactions over time often does not consider the effect that prior trial rounds can have on subsequent ones, and, thus, in the formation of trust.

We analyze participants' trust development in AI advice in a sequential legal decision-making task. We measure trust in two ways: as behavior, which is defined by the extent to which people change their initial assessment in the direction of AI advice ("reliance"), and as people's beliefs or cognitive assessment of the AI, which we capture as self-reported answers ("trust"). Building on previous work [59, 76], we investigate how different model accuracies and system explanations affect trust and reliance. In line with existing literature and common sense, we assume that low-accuracy AI advice will not be able to build up the same level of trust and reliance as high-accuracy AI advice. And we expect that people will be able to learn to trust in and rely on higher accuracy systems more, given enough time to experience the difference. Second, we suggest that human-like explanations of AI advice are more likely to increase trust in and reliance on AI advice than abstract explanations. When considering these effects, we account for the temporal dimension of experiencing AI advice: evaluations of prior trial rounds may affect trust and reliance evaluations for subsequent trials. We investigate this matter in two ways – we analyze to what extent a given trial impacts a subsequent one (lagged trust/reliance) and to what extent the aggregate of prior trials impacts a subsequent trial (aggregate trust/reliance). Most studies on trust in HAI disregard the fact that people may observe and update their thoughts and behavior over time, even when interacting over multiple rounds in a relatively short time span (as in our experiment). As the interaction with an AI model progresses, people's beliefs and behavior may be based on prior interactions rather than on contextual circumstances such as explanations or accuracy alone. Therefore, accounting for prior interactions may contribute to a better understanding of trust beliefs and reliance behavior concerning interactions with AI models.

With our study, we hope to contribute new insights to the current HAI literature by extending existing work in several ways. We do this by examining trust and reliance in AI advice in an applied and rather complex decision problem (instead of a simple classification task) across the sequence of 20 trials (as opposed to single-shot tasks), applying multiple trust measures (trust, reliance, post-trial trust), and examining the effect of prior interactions on subsequent trials. The following research questions guide our work:

- **RQ1:** Do trust and reliance vary based on different model accuracies and types of explanations?
- **RQ2:** How do trust and reliance develop over time for different model accuracies and different output explanations?
- **RQ3a:** To what extent do prior trust and reliance evaluations in AI advice impact people's trust and reliance evaluations for subsequent AI advice in a sequential task?
- **RQ3b:** How do trust and reliance impact each other over the course of HAI?

2 RELATED WORK

Having trust in a well-functioning AI model that can improve human-decision making is an essential prerequisite for thriving HAI. A well-known definition of trust comes from Mayer et al. [66] who define trust as the "willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" (p. 712). A second understanding of trust emphasizes a successful outcome based on trusting AI (advice), which is defined as as "the attitude that an [AI] agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability" [57]. Subtle differences in definitions exist. For example, trust can be defined as a personal trait or disposition: some people are inherently more trustworthy than others independent of the context [42]. In contrast to the character trait trust, (cognitive) trust beliefs can change over time as information about the trusted system becomes available [82]. Trust can also be expressed and measured as behavior: in a decision-making scenario with AI support, people can express trust in the system by the extent to which they are willing to accept AI advice [56]. Depending on the scenario, this may imply that a person overrides their own decision to comply with the AI fully, or it may imply that a person changes their own decision in the direction of the AI recommendation. Both can be interpreted as trusting behavior.

In our study, we measure trust in two ways by distinguishing between trust beliefs (which we denote as **TRUST** in the following) and trust behavior (which we denote as **RELIANCE** in the following). Previous work has shown that the two dimensions correlate with each other [56], but measuring behavior and beliefs could also lead to different trust results: people may state that they trust but still not act on it, or vice versa. Trust beliefs underlie a cognitive, affective process and may be formed over time, based on factors like experiences, inherent beliefs, or trust dispositions, and can thus be described as relatively stable and rather resistant to change[42]. In contrast, we argue that trust behavior may be less stable. For instance, people may consider specific circumstances when relying on AI advice, for example, a specific task at hand, other external factors, such as the risk of a situation, which demands more flexibility in trust behavior [22, 38, 82]. We furthermore argue that people could even vary trust in AI advice during a specific sequential task, where they are familiar with some tasks but lack knowledge for others, where AI advice would be beneficial. From this, we conclude that trust beliefs appear more stable over time, while trust behavior appears more dynamic, as they are based on case-by-case assessments of (individual) tasks. Likewise, we argue that the two dimensions interact: (prior) trust beliefs can influence trust behavior. For example, a positive experience with accurate AI advice will strengthen trust beliefs and eventually influence reliance on AI positively, similar to previous work [11, 54].

The scientific community has already gained considerable insights into the processes and factors that influence trust and reliance on AI systems. On the one hand, system properties can impact user trust: system transparency [12, 15, 16, 69] as well high accuracy levels of AI positively influence user trust and reliance [4, 27, 33, 65, 96], as well as visual system properties, such as human-like representations [49, 50]. Communication possibilities between the system and its users are another factor influencing trust and reliance positively. For example, explaining system errors helps to protect trust in AI [29]. This can be in the form of providing justifications or apologies after a mistake occurred [30, 55]. Verbal explanations increase the willingness to follow AI advice more compared to visual cues [91]. Generally, trust and reliance are higher for human-in-the-loop scenarios compared to those where the user has little or no control The same accounts for human-in-the-loop setups [1, 24, 35]. Furthermore, trust and reliance in AI are impacted differently based on the user and their skills and traits: self-confidence in oneself affects appropriate trust in AI negatively [18], (domain) expertise leads users to trust AI advice too little [61], and (political) conservatism is associated with low comfort in AI, and, hence, results in lower levels of trust in AI systems [13]. Finally, trust and reliance on AI advice are impacted by the context of a decision-making situation: certain emotional states mediate trust in AI [31], uncertainty and complexity hinder trust in AI [26], and time pressure can lead to an inappropriate level of trust in AI [88].

In conclusion, in many decision scenarios, people accepting AI advice has led to better outcomes [60, 61, 79]. Then again, rather than make people trust AI advice as much as possible, we should strive for appropriate levels of trust: people should neither place too much trust in a sub-optimal AI system nor put too little trust in a well-performing one.

2.1 Trust and Reliance Development Over Time

Although trust research regarding HAI gains recognition, it is often discussed without a longer-term or repeated-scenario perspective [43, 90]. This is crucial though, given that it is likely that when people familiarize themselves with and understand a system better, trust and reliance will likely alter from being fragile and unsteady to more stable [67, 68]. Research offers no clear answers as to whether trust in AI decreases or increases over time and on which conditions this depends. Yang et al. [95] found that self-reported trust increases with automation successes but decreases even after AI failures. Similar studies from Chacon et al. [14], and Nourani et al. [72] show a sharp decline in reliance after early AI errors, which was not recovering to the same levels when experiencing good AI performance afterward. Nourani and colleagues and Desai et al. [25] assume a primary-recency effect: initial (and late) interactions affect trust and reliance the most when measured after exposure to the system. At the same time, some comparable studies show that trust grows over time: Chiou et al. [17] found that study participants indicated trust in an intelligent (robotic) system as operators understand how to work with it throughout several interactions successfully. Similarly, Manchon et al. [64] assessed trust in an automated driving system over time (three assessments in four months) and found that trust increased over time for both trustful and distrustful drivers. Manchon et al. posit that several positive interactions in the early phase of the study supported even distrustful participants in gaining trust quickly. From this, we conclude that initial trust is crucial for more extended interactions and efforts to motivate people to start using automated systems in the first place [19, 64]. Tolmeijer et al. [90] add that initial trust is important to accept better later mistakes, strengthening long-term interactions. Yu et al. [97] studied trust dynamics based on different model accuracies. Participants solved a binary quality control task (assessing whether drinking glasses were produced correctly or not). Their findings suggest that trust trajectories differ based on accuracy levels: trust increased over time when model accuracy was 80% or higher. However, it decreased with 70% accuracy. They also showed that AI failures at different time stages during the interaction demonstrated different implications for the trust change: participants who had time to become familiar with the system did not experience a decrease in trust after AI errors. Trust stabilized at the end of each task block. Over time, study participants formed a stable mental model for themselves, also described as the inertia of trust. Overall, they propose a phase model in which people learn to what extent a system can be trusted (Phase 1), adjust this learned trust (Phase 2), and finally fine-tune it (Phase 3). Thus, people learn step-by-step to rely on a system once they find it trustworthy enough [77].

Furthermore, we contend that it is important to build on existing trust studies such as that of Yu et al. [97] and understand how people adopt AI advice during an interaction for a longer period of time or for repeated tasks. So far, empirical work of the latter is limited. Falcone & Castelfranchi's theoretical model [32] challenged the basic notion that positive AI advice will always lead to an increase in trust and argued to consider the cognitive process of people updating their beliefs. Accordingly, Yang et al. [95] found that people adjusted trust based on their moment-to-moment interaction with an automated system, which increased with successful advice and decreased after system failures. Hafizoglu & Sen [41] measured human-virtual agent collaboration based on a repeated trust game (five tasks, where trust was measured three times with a three-item scale) and found that positive past experiences affected future ones but also that previous experiences were able to overwrite initial trust levels.

In summary, earlier research on trust development in AI suggests that the initial phase of HAI significantly determines how trust in a system develops and that trust stabilizes after people become familiar with AI [17, 97].

Moreover, positive AI impressions strengthen trust, but negative AI impressions result in loss of trust, which can be challenging to restore [25, 90]. We argue for examining HAI's longer or repeated interaction sequences in more detail to see how individual interactions affect trust over time. Thus, we follow up on the cited work by measuring both trust and reliance over the sequence of 20 tasks where study participants receive AI advice. Different from Hafizoglu & Sen [41], our scenario aims to augment human decision-making and does not aim for mutual collaboration between the two parties. Furthermore, we extend research on trust development by considering people's prior AI experiences during a sequential interaction and how they influence future beliefs and behavior. Including the latter in our study, not only could we learn about the impact of different system properties (explanations, accuracy) on trust and reliance, but we could also better control the potential influence of prior decisions to trust and rely on AI advice.

2.2 Model Accuracy & Failure

In many (though not all) prediction tasks, AI tools can match or outperform their human counterparts [36, 40]. Model accuracy is an essential criterion for whether an AI tool is considered helpful and a prominent determinant of trust [85]. Despite its capabilities, systems can fail at certain points, and people must remain alert to detect both generally faulty systems and occasional wrong decisions from decent systems. Studies show that consistently low accuracy is indeed observed and acted upon: Yin et al. [96] found that trust was significantly affected by the actual level of accuracy of a system; in comparison, the effect of stated accuracy only has a negligible impact on trust. Similarly, information on how reliable a model is increased trust and performance [33]. Moreover, trust was (lastingly) damaged and recovered only slowly after experiencing errors in performance [27, 29, 65, 97]. Yu et al. [97] manipulated model accuracy on four levels (70%, 80%, 90%, and 100%), giving false positive/negative advice accordingly: trust decreased only in the 70% condition in comparison to the other tested accuracy levels. In line, Papenmeier et al. [76] compared trust levels of participants that interacted with a high, medium, and low ("antagonistic") accuracy and found that participants showed adequate levels of trust.

To summarize, people can distinguish between good and poor AI advice based on its accuracy, which is a crucial requirement for an appropriate level of AI reliance, as it is important that people only adopt advice from models that are more accurate than they are themselves. We will apply different levels of accuracy to compare how trust and reliance develop. Like Yu or Papenmeier, model accuracy is not explicitly stated and can only be anticipated by the advice itself. Our goal is to study to what extent participants can distinguish between higher and lower-performing models over time and, more so, to what extent this is reflected in the development of trust and reliance levels over the 20-trial sequence.

2.3 Model Explanations

With increasing computing power, AI systems exhibit increasing levels of complexity and, consequently, system opacity. Recognizing the imperative for transparency and the ability of people to comprehend AI and somehow maintain control over it, there has been a growing emphasis on explaining AI output, which made the research domain of explainable AI (XAI) a key one of current AI research. Studies such as those by Arrieta et al. [2], Lim et al. [59] and Ribeiro et al. [80] show the central role of explanations in improving our understanding of AI systems and how they work, a key prerequisite for the acceptance of AI advice.

Explanations assist in understanding the predictions of a model better. Those can take different forms; for example, a prediction can be accompanied by a text-based *how, when, why (not)* explanation [47], they can be supported by visuals [20, 45], (counterfactual) examples [39] or numerical values (e.g., confidence levels) [98]. Furthermore, they can be designed as interventions (warnings, nudges) or post-decision arguments (apologies, promises, justifications) [9, 10, 29, 30, 50, 53, 73, 84, 91]. Explanations (in comparison to no explanations) have

also been shown to impact trust (over time) positively [43, 69]. For example, Lim and Dey [58] found that understanding and trust in a system are highest with model explanations.

Although most studies found positive effects of AI explanations on trust and reliance, some studies observed feelings of manipulation by AI [10] or overconfidence in AI [93]. Papenmeier et al. [75, 76] find evidence that not all explanations are helpful, and some might even be harmful: they discovered that adding nonsensical or random explanations hurt trust (as one would hope). Furthermore, they find that explanations do not improve trust when people interact with a sufficiently accurate system; they argue that the type of explanations that were used, highlighting words in a text document, did not improve decision-making as it did not add any value for participants because these explanations were of statistical rather than causal nature, which only partially supports human understanding. Finally, Nourani et al. [71] tested whether trust differs for meaningful versus meaningless explanations. They defined meaningfulness as the level to which explanations were perceived as meaningful in the human context. Their results show that participants significantly underestimated model accuracy when providing weak (less human-meaningful) explanations, as they did not understand the results. The study, which was conducted with non-expert participants, claims the need for "human-interpretable" explanations. Those findings are somehow supported by Tocchetti and Brambilla [89] who argue that AI explanations must be more approachable for humans and propose ways to improve understandability in their review. In conclusion, we hypothesize that explanations offer added value if they align with human argumentations or are human-interpretable. Thus, we assume that explanations that are expressed in a rather human-like and contextual way could lead to increased trust and reliance on AI advice compared to abstract (keyword) explanations that offer neither additional argumentation nor are rich in form or expressiveness.

2.4 People's Perception & Traits

Factors that can furthermore impact trust and reliance on AI advice are user traits. For one, the level of (domain) expertise and confidence of a person can interfere with correct trust assessments: previous work shows that lay people over-rely on AI advice when experiencing correct outputs (especially at the beginning of an interaction), whereas expert users rely on AI advice less, even after their own performance decreased along a task [61, 72]. We account for domain expertise as a covariate in our analysis to see how legal expertise affects trust and reliance on AI advice. Specifically, we assume that participants with a higher level of legal expertise have less trust in AI advice and, therefore, rely on it to a lesser extent. Another trait that affects trust in AI is a person's affinity to engage with technology, which positively affects collaboration with AI systems [44, 70]. Affinity for (new) technology precedes the notion that people can somehow create a better mental model of what systems can do: by "opening" the black box, they are more willing to trust a well-performing system. We included affinity to technology as another covariate based on our expectation that participants with a higher affinity for technology show equivalently higher scores for trust and reliance on AI.

3 STUDY DESIGN & METHODOLOGY

We analyze the development of trust and reliance using a decision-making task based on real criminal law data, for which participants are supported by AI advice. Our study follows a 2x2 between-subjects design. Participants are randomly assigned to one of the four groups, in which we manipulate model accuracy (high vs. low) and explanation type (human-like vs. abstract). They are asked to estimate jail time for 20 legal cases, for which they receive support through numerical jail time estimates which are supported by textual explanations (see Figure 1). All 20 criminal cases were selected from the Dutch database *de Rechtspraak* [23] so that all actual case verdicts (= ground truth) were known.

This paper discusses and compares the findings of two studies that are both based on the legal decision-making task as described above and further clarified in Figure 2. **Study 1** is the "original" experiment [46]. **Study 2** is

a conceptual replication of our original experiment, which follows essentially the same study design but with several improvements. For example, we tested slightly different accuracy levels (two levels) and excluded the explanation type manipulation. Furthermore, we fully randomized the AI advice and extended the post-trial trust measurement. We describe all changes in detail in Chapter 5.1. Although our two studies are almost identical, we mention for clarification that Chapters 3 and 4 will be concerned with **Study 1**, and the replication study (**Study 2**) separately in Chapter 5.

3.1 Target Variables

We define trust based on two dimensions: trust beliefs and trust behavior. Although it is known that trust beliefs and trust behaviors correlate [56], they are not identical. We also consider the possibility that trust beliefs influence reliance behavior and vice versa: for example, we argue that positive interactions can reinforce future willingness to follow AI advice. To what extent do trust and reliance influence each other? This will be part of our analyses.

3.1.1 Trust (beliefs / cognitive trust). We measure trust beliefs by asking participants to indicate their level of trust repeatedly after each legal case, which we denote as TRUST in our study. Subjective measures are used to capture inherently subjective trust data, reflecting an individual's perspective [81]. These psychological constructs are usually measured with survey scales. We do this with a single-question item asking participants to indicate their momentary level of trust in the AI system (1 = no trust at all, 10 = full trust). By using a simple measurement to capture trust development over time, we avoid overburdening participants with too many questions, similar to the study setup from Yu et. al. [97].

3.1.2 Reliance (behavioral trust). We measure behavioral trust as the willingness of participants to rely on AI advice, which we denote as RELIANCE in our study. We, therefore, quantify the extent to which participants comply with the AI advice. We measure the willingness to comply with AI advice based on the concept of Weight on Advice (WoA), which is calculated by weighing two consecutive estimations of the participant, their first estimate before seeing the AI advice and the second estimate after seeing the AI estimate [87]. In theory, the continuous outcome ranges from 0 (= people completely ignore the AI estimate and stay with their own estimate) to 1 (= people completely rely on the AI advice and adopt it as their own estimate). In practice, these given margins are often exceeded as people compensate for AI advice by overshooting or undershooting the AI estimate to adjust their own decisions. Applying this measurement, we follow other studies in the trust in AI literature [8, 61].

$$\text{Weight on Advice} = \frac{\text{Second Estimate} - \text{First Estimate}}{\text{AI Advice} - \text{First Estimate}}$$

Scharowski et al. [81] point out that one reason the overall results of the AI literature on trust are inconclusive is that there are no standardized measures, partly because definitions of trust are ambiguous. We also see that most studies focus on either trust beliefs or trust behavior. Expecting to capture a potential causal relationship (correlations or interactions) between trust and reliance, we include both. Measuring reliance may come with the advantage of quantifying trust somewhat objectively. Yet, a potential pitfall could be that not being close to the AI estimate does not necessarily mean that people do not rely on it. Still, it could also mean that participants are confident in solving a task independently. If the user's initial estimate and the AI estimate are close from the beginning, the deviation is naturally small. Measuring trust as a self-reported dimension can help capture an immediate psychological response based on direct experiences with AI. We assume that trust beliefs stabilize over the course of time when participants have sufficient time to determine to what extent they can rely on its output. However, trust may also be influenced by a personal inclination that would not reflect the attitude towards the AI advice per se but may indicate a person's trust attitude in general.

3.2 Predictor Variables & Study Manipulations

3.2.1 Model Accuracy. The first main predictor variable in our study is model accuracy. It is executed as a two-factor variable with high and low model accuracy. We define accuracy as the extent to which the AI advice deviates away from the correct jail time (= ground truth). As we wanted the deviation error to occur as naturally as possible, we calculated the AI estimates by adding a normal-distributed error around the actual jail time: the high-accuracy model deviated +/- 10% (95% interval) from the correct jail time, and the low accuracy model deviated +/- 50% (95% interval). All estimates were weighed against each other to avoid systematic bias, such that AI estimates were either higher or lower than actual jail time. Overall, we wanted participants to perceive the accurate model as more competent than the inaccurate model and, therefore, follow the former one more: even though we did not state model accuracy or resolve the correct results of participant versus AI per case, participants were able to compare their own decisions to the AI advice and would experience how accurate the AI was over the sequence of 20 trial rounds. Comparing the AI estimates of both models, the high-accuracy model is substantially more accurate with a mean absolute difference of 3.2 months (an average 10.1% deviation from the correct jail time) versus the low-accuracy model with a mean absolute difference of 16.5 months (an average 46.2% deviation from the correct jail time). Still, there were a few cases where the low-accuracy model performed equally well or even better than the high-accuracy model (see Table 1) due to chance. Unlike Yu et al. [97], we applied (close-to-)continuous measurements, represented by AI estimates of jail time in months, instead of a binary choice task and only manipulated two different accuracy levels. All estimates were pre-calculated and coded to be the same for all participants (Table 1).

In our regression models, we refer to our accuracy manipulation with the name **HighAccuracy** using indicator coding, where value 1 represents the high accuracy and 0 represents the low accuracy condition.

3.2.2 Model Explanations. The second main predictor in our study is the explanation type. It is executed as a two-factor variable, human-like and abstract explanations. Our experimental setup was inspired by important work in the field of XAI, such as Ribeiro et al. [80], who propose LIME as an explanation technique. Similar to the latter, where AI predictions are based on textual information that supports an AI decision, we prepared the explanations for each legal case by identifying a set of case-specific keywords. To do this, we read the original cases, on the basis of which we created a shorter version for each criminal law case. Next, we picked keywords that best represented the cases. The selection of case-specific keywords was then applied to two different types of explanations: human-like explanations were constructed as complete sentences with contextual arguments intended to evoke a sense of human-like intelligence. For the second explanation type, abstract explanations, we used identical keywords but presented them as a simple series of words. The design of our explanations could be compared to highlighting specific keywords in a text, similar to the approach of Ribeiro et al. [80]. Figure 1 demonstrates our two explanation types based on an exemplary legal case.

Following the AI design proposals of Knijenburg and Willemsen [51] and Nourani et al. [71], we argue that displaying full-sentenced explanations elicits comprehensibility and allows the feeling of interacting with a rather intelligent, human-like system. Following this line of thought, we highlighted the different (explanation) capabilities by introducing the two systems to the study participants with different system names (human-like: *AI legal case analysis system*, abstract: *jail time calculator program*) and providing details to the inner workings of each. As described above, the explanations and supporting cues could lead people to perceive the human-like system as more capable and potentially reliable, thus enhancing trust and reliance on its recommendations.

In our regression models, we refer to our explanation type manipulation with the label **HumanExplanation** using an indicator coding, where the value 1 represents the human-like explanation type and 0 the abstract explanation type.

3.2.3 Lagged Trust & Reliance. In addition to our main predictor variables, we want to explore the impact of prior trust and reliance experiences that participants gain during our 20-trial sequence on trust and reliance for subsequent trial rounds. We define two separate sets of lagged predictor variables and include them as covariates in our analysis. The first variable set is defined as the effect of immediate prior trust and reliance interactions on a subsequent trial round. As an example, we assume a potential effect of trust and reliance from trial round 3 onto trust and reliance evaluation in trial round 4. The second variable set is as the effect of aggregate prior trust and reliance on a subsequent trial round. Different from the immediate interactions, we accumulate the trust and reliance experience of all past trial rounds up to the subsequent trial round. For example, we assume that the accumulated trust and reliance experiences from trial rounds 1 up to 12 may affect how participants assess the AI advice in trial round 13. We include the two lagged variables, assuming they may help us better explain trust and reliance, based on the assumption that model accuracy and explanation type alone may not fully explain potential effects.

In our regression models, we refer to our lagged variables with the names **ImmediatePriorReliance**, **ImmediatePriorTrust, AggregatePriorReliance** and **AggregatePriorTrust**.

3.2.4 Person-specific Traits. Finally, we measure several person-specific covariates and their influence on our target variables. We ask participants to indicate their level of legal expertise (10-point Likert scale: 1 = no expertise at all; 10 = very high). Logg et al. [61] showed that expertise could lead people to over-trust their own skills and, therefore, to mistrust or ignore an intelligent system's recommendations. Following the assumption that personality traits affect reliance or trust, c.f. [83], we propose to measure whether prosocialness influences trust in AI. We apply the Prosocial Behavioral Intentions Scale [5] (α = .81). Participants rate their willingness to get involved in social situations on a 7-point Likert scale (1 = I would definitely not do this; 7 = I would definitely do this). Finally, we ask study participants to indicate their affinity for technology, using the 4-item Affinity for Technology Interaction Short Scale (ATI-S, α = .87) from Wessel et al. [94]. Participants indicate to what extent they agree with the four statements on a 6-point Likert scale (1 = completely disagree; 6 = completely agree).

In our regression models, we refer to our person-specific variables as **LegalExpertise, Age, TechAffinity**. For the gender of our participants, we use the indicator variable name **Female** (female participants are described with a value of 1 and males with a value of 0).

3.3 Procedure

At the beginning of the experiment, participants were randomly allocated to one of the four groups. After being introduced to the terms of the experiment and accepting the consent form, participants were introduced to the support system, which was either the system with human-like explanations, called *AI Legal Case Analysis System*, or the system with simple explanations, called *Basic Jail Time Calculator program.* To support active engagement with the system, we asked participants to confirm that they had read the introduction carefully by ticking the corresponding box. In the second introduction part, participants learned about the task procedure (Figure 2): Participants read the case and indicated their initial jail time estimate (1). After seeing the system calculate a result (interactive visual element) (2), they were automatically directed to the AI output, the numeric estimate, and the explanation (3). Participants then adjusted or confirmed their second estimate (4). They learned about the correct verdict of the case (5) and finally indicated their momentary trust in the AI system with a slider going from 1 = no trust at all to 10 = full trust (6). The task procedure was repeated for 20 legal cases. To avoid order effects, cases were presented in random order. The last part of the study covered the following topics as a questionnaire: perceived level of intelligence of the AI system, perceived level of accuracy of the AI system, participant's relationship towards technological systems (affinity to technology), their willingness to participate in social situations (prosocialness), their level of legal expertise, and their demographics (age, biological sex). The study closed with the debriefing and the remuneration of participants. Participants took approximately 20

Legal Case	Actual Jail Time	AI Estimate (High)	AI Estimate (Low)
1. Street Stabbing	24	21	10
2. Partner Murder	96	102	24
3. Possession Illegal Fireworks	6	7	1
4. Cocaine Import	10	10	13
5. Theft Retirement Home	6	5	8
6. Money Laundering, Drug Possession	30	32	17
7. Online Scam	4	5	4
8. Violence Against Police	12	12	12
9. Attempted Family Murder	72	81	30
10. Child Death	36	43	56
11. Theft, Assault, Drug Possession	6	5	11
12. Attempted Murder	48	52	85
13. Home Burglary	7	7	10
14. Attempted Murder, arson	84	75	17
15. Premeditated Murder	120	134	126
16. Fraud	12	12	12
17. Weapon Possession	12	13	23
18. Physical Partner Abuse	5	4	3
19. Gun Possession, Money Laundering	30	28	22
20. Weapon Possession, Drug Possession	48	50	29

Table 1. Overview of the criminal law cases used in the experiment. "Actual Jail Time" and (high/low) "AI Estimate" are given in months. Based on the actual jail time (= ground truth), AI estimates were HIGH in accuracy (+/- 10 percent margin from ground truth) or LOW in accuracy (+/- 50 percent margin from ground truth). Estimates were hard-coded before the experiments, and every participant received the same estimates depending on the experimental condition.

minutes to finish. We used identical AI estimates for all participants: depending on the factors (accuracy, type of explanation) in the four conditions, all estimates and explanations were pre-formulated and pre-calculated (see Table 1) without using an actual AI system. Even though this was part of the study's cover story, the intelligent systems were purely fictitious and did not perform any actual calculations.

3.4 Participants

We approximate our sample size with an a priori power analysis (ANOVA fixed effects, main effects, and interactions, effect size f = 0.25, α error = 0.05, power = 0.9) based on our manipulation factors (model accuracy: two levels; model explanations: two levels). This resulted in a total sample size of N=171 participants [34], which can be interpreted as an approximation of the necessary sample size under the assumption that there is a low correlation between the repeated measurements. We base our calculation on similar studies [75, 76, 90, 97] with repeated trials with medium effect sizes (0.25) and comparable sample sizes. We recruited participants via the Prolific research platform [78] based on participants' age (18 years and older) and current location (UK). Additionally, we selected participants with expertise in the field of law to increase the variance of domain expertise for our sample. Unfortunately, this minimized the available sample to <500 eligible participants, which led us to recruit only half of our needed participants based on the selection criteria (84 of 171 participants). We recruited the remaining 87 participants without the criteria of a (professional) background in the legal sector. In addition to

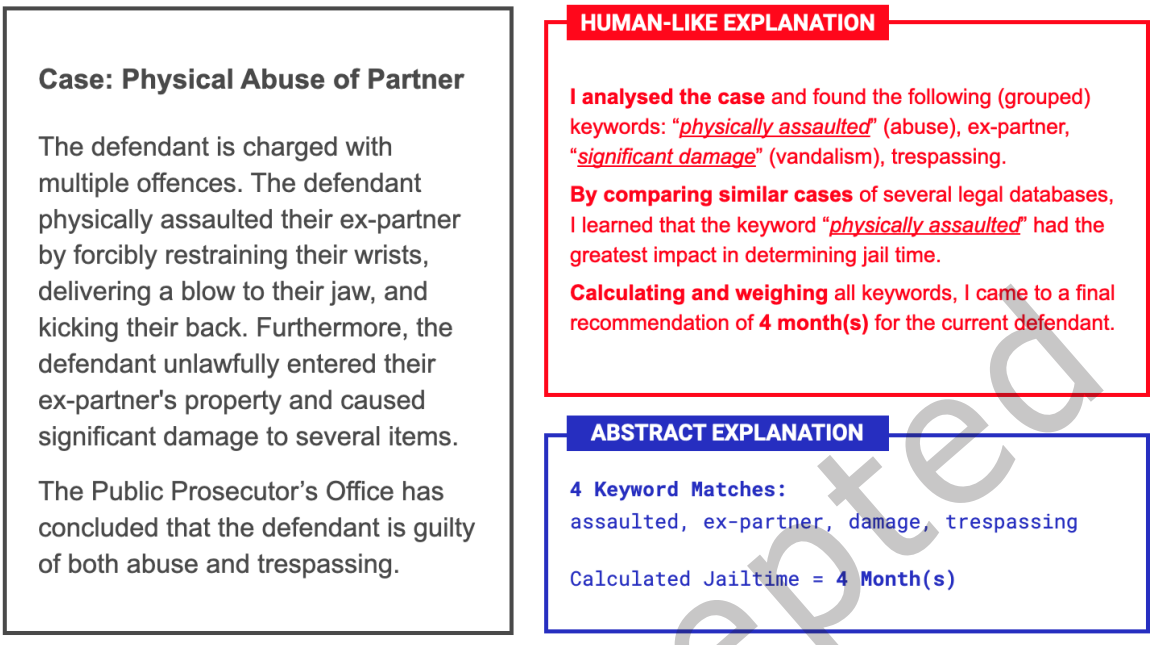

Fig. 1. Exemplary legal case: Descriptive text (black) with the human-like (red) and abstract (blue) explanations which are shown in step two of each procedure

our required sample, we recruited N=10 additional participants to pilot our study. Each participant received, on average, £5.87 as compensation. Both the pre-test and study were conducted between May 24 and June 1, 2022.

N=204 participants registered for the study from which N=171 participants (84%) fully completed it. N=23 participants were excluded because of age restrictions (younger than 18 years), not finishing the full study, or showing straight-liner answer behavior for several question inventories. We also excluded the pilot study participants from our final data set. 71% of participants were female, they were 36.6 years old on average (Min: 19, Max: 69, SD: 12.33), their mean legal expertise score was 4.2 (Min: 1, Max: 10, SD: 2.51), their mean prosocial level was 6.0 (Min: 2, Max: 7, SD: 0.87), and their average tech affinity score was 3.7 (Min: 2, Max: 5, SD: 0.54).

3.5 Hypotheses

Our study aims to measure the development of trust and reliance based on four different study conditions. In line with previous studies that tested the effect of model accuracy and model explanations, we expect that trust and reliance will be higher in the high-accuracy (vs low-accuracy) conditions. We furthermore believe that trust and reliance increase more (vs less) with human-like explanations. We, therefore, assume the following hypotheses, which were pre-registered under the Open Science Framework [1].

- **H1:** Trust and Reliance are higher for high model accuracy than for low accuracy model.
- **H2:** Trust and Reliance increase more / decrease less over time with high model accuracy than with low model accuracy ("high accuracy protects trust and reliance better").
- **H3:** Trust and Reliance are higher with human-like explanations than abstract ones.

[1] https://osf.io/avux5/?view_only=69230d1656b14f8aaf9a6f34030bef7b

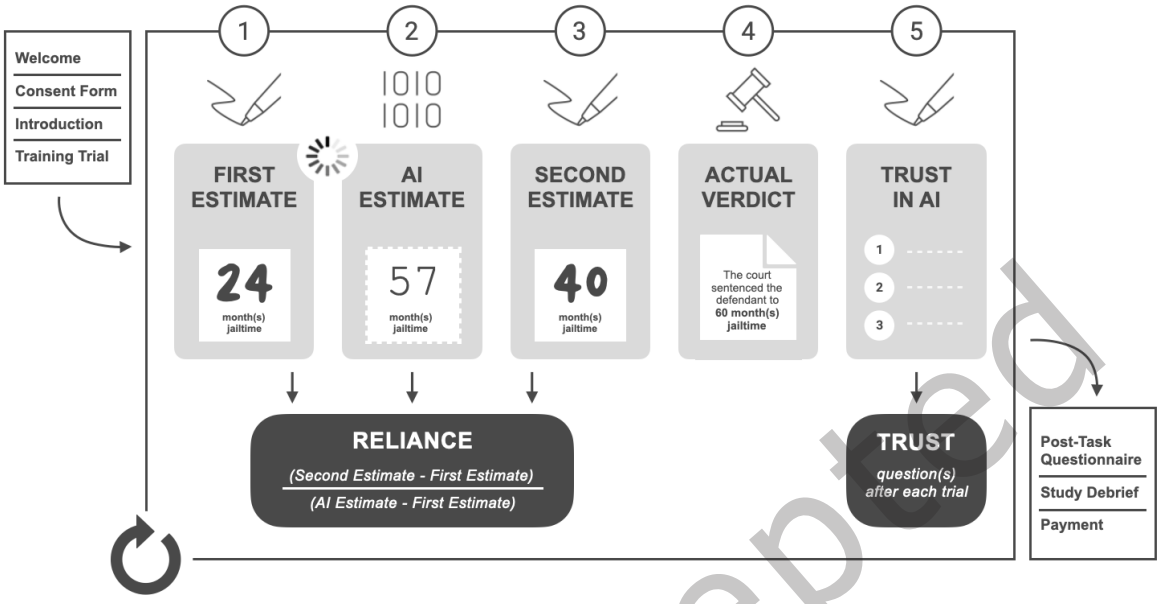

Fig. 2. Study Procedure: After the introduction (consent form, introduction, training trial), participants solve a sequence of 20 legal tasks trials. Study participants finish the study with a set of post-trial questions and the study debrief before they are guided back to Prolific for reimbursement. We note that the post-trial trust measurement differed for Study 1 (original), where we used one question with a 10-point Likert scale, and Study 2 (replication), where we used three question items with a 7-point Likert scale. Find more details regarding the different setups of both studies in Chapter 5.1.

- **H4:** Trust and Reliance increase more / decrease less over time with human-like explanations than with abstract explanations ("human-like explanations protect trust and reliance better").

In addition to our pre-registered hypotheses, we explore the relation of the covariates of legal expertise, tech affinity, age, and gender on trust and reliance.

In addition, we are interested in the potential correlations between our target variables, trust and reliance. Not many experiments in HAI apply different trust and reliance measurements in one experiment, so our analyses could add valuable insights into how beliefs and behaviors affect each other.

Finally, we analyze whether people's trust and reliance assessments are influenced by prior interactions with AI advice. As far as we know, these additional analyses on the effect of prior trials on subsequent ones have not yet been applied in a context such as ours. Our analyses are exploratory in nature and we have not pre-registered any hypotheses.

4 RESULTS

4.1 Statistical Analysis

Our data consists of multiple measurements per participant, making it likely that standard multiple regression models are not appropriate. To take this "clustered" or "nested" (cases nested within participants) structure into account, we use multi-level regression. In this standard approach for data with a nested structure, we account for

participant heterogeneity by including a random intercept. Alternative names for essentially the same procedure are "repeated measures ancova" or "mixed models". We distinguish between models with either trust or reliance as the target variable and between models for (Study 1) and (Study 2). We discuss the effect of model accuracy and explanation type on trust and reliance based on several multi-level regression models for all four of these options. We start with a baseline model with just the model accuracy, the explanation type, and the trial (round) as predictors. We then extend these models with additional covariates, such as interactions between the trial (round) and the conditions, some participant characteristics, and assessments in earlier rounds. We use this approach to be able to assess the extent to which the estimated coefficients of our main predictors (model accuracy and explanation type) are sensitive to the kind and number of other predictors included in the model. The analysis for the original study (Study 1) can be found in Tables 3 and 4.

From the initial 3,420 data points for our target variable **Reliance** (calculated based on Weight on Advice, see Chapter 3.1), we excluded 253 observations where participants' initial estimate was identical to the AI estimate (in which case there is nothing to infer about the extent to which AD advice was followed). We excluded 513 further observations due to our defined Weight-of-Advice (WoA) margins of 0 and 1, which is the usual proportion choice in most WoA application studies [3]. Our final set of observations for reliance was 2,654 decisions. For our target variable **Trust**, which was defined by a single-item question at the end of every legal case trial, we derived 2,907 decisions (after 513 observations were excluded in line with our WoA procedure).

As our main predictor variables, we included explanation type, referred to as **HumanExplanation**[2], and model accuracy, referred to as **HighAccuracy**[3]. We included the variable **Trial** which defines the sequence of 20 legal case trials and allows us to measure trust and reliance over time. We furthermore included the interaction variables **HumanExplanation*HighAccuracy, Trial*HighAccuracy** and **Trial*HumanExplanation**. Furthermore, we included the following covariates in our models: **Age**, Gender, referred to as **Female**[4], **TechAffinity** and **LegalExpertise**. Finally, we included two lagged variables to test the effect of participants' prior trust and reliance experiences of (a) trial round(s) on a subsequent trial round. **ImmediatePriorReliance** and **ImmediatePriorTrust** indicate any effect of trust and reliance of the prior trial round on a subsequent trial round (for example, the influence of trust and reliance experiences of trial round 2 on trial round 3). Moreover, we included two predictor variables to account for aggregate prior experiences concerning trust and reliance, which we refer to as **AggregatePriorReliance** and **AggregatePriorTrust**. These measure the accumulated trust and reliance experience up to the point of the subsequent trial. For example, the influence of trust and reliance experiences of trials 1 to 19 on trial round 20.

4.2 People perceive the AI model differently concerning intelligence & accuracy

To see whether our experimental conditions affected participants, we asked participants to rate the AI system on system intelligence (assuming that participants would rate the AI model giving human-like explanations higher) and model accuracy on a 10-point Likert scale. We performed a two-sided t-test (equal variance) for both questions. Participants perceived the model with human-like explanations ($M = 6.39$, $SD = 2.45$) as significantly more intelligent than the model with abstract explanations ($M = 5.64$, $SD = 2.35$), $t(169) = -2.0$, $p < 0.04$. Participants also perceived the high-accuracy model ($M = 7.39$, $SD = 1.67$) to be significantly more accurate than the low-accuracy model ($M = 4.28$, $SD = 1.84$), $t(169) = -11.55$, $p < 0.001$. These significant results suggest that the study manipulations worked.

[2]Explanation Type is indicator-coded: human-like explanations are coded as 1 and abstract explanation as 0.
[3]Model Accuracy is also indicator-coded: high accuracy is defined with value 1 and low accuracy with value 0
[4]Gender is indicator-coded: female participants are defined with value 1 and male participants with value 0

Reliance:			**Trust:**		
	High Accuracy	*Low Accuracy*		*High Accuracy*	*Low Accuracy*
Human Explanations	M = 0.61 (0.01)	M = 0.31 (0.10)	*Human Explanations*	M = 6.80 (0.07)	M = 4.13 (0.08)
Abstract Explanations	M = 0.49 (0.49)	M = 0.32 (0.10)	*Abstract Explanations*	M = 6.24 (0.10)	M = 4.14 (0.08)

Table 2. Comparing mean levels of trust and reliance in the four experimental groups, based on the two conditions Model Accuracy (High, Low) and Explanation Type (Human, Abstract). Standard Errors are given in round brackets (95% confidence interval).

4.3 Trust and reliance are higher for high-accuracy models (H1, H2)

We compare the mean levels of trust and reliance (across trials) per experimental groups (see also Table 2). Reliance shows higher levels of high accuracy across both explanation types (human-like: M = 0.61, abstract: M = 0.49) compared to low accuracy (human-like: M = 0.31, abstract: M = 0.32). The same can be observed with trust, where mean levels are higher for high accuracy (human-like: M = 6.80, abstract: M = 6.24) than for low accuracy (human-like: M = 4.13, abstract: M = 4.14). We test for the significance of these differences in the regression models.

Models 1 (R^2 = 0.12) and 4 (R^2 = 0.20) (Table 3) show the main effects of model accuracy and explanation type and models the effect of trust and reliance over time. Consistent with H1, we find that trust and reliance are significantly higher with **HighAccuracy**. Across the conditions, reliance is stable over time (no effect of **Trial**, β = −0.001, p = 0.230) but trust seems to develop over time (significant effect of **Trial**, β = 0.019, p = 0.001). Model 2 (R^2 = 0.13) and Model 5 (R^2 = 0.21) allow us to see whether the effect of accuracy is different over trial rounds by including the interaction of **Trial*HighAccuracy**. For reliance, we see that trial (baseline model is for low accuracy) becomes negative (Model 2: β = − 0.004, p = 0.018) but the interaction of trial with high accuracy is significantly positive (Model 2: β = 0.007, p < 0.001). This reflects the results in Figure 3 with a downward slope for the low accuracy condition for reliance. Trust (Model 5) now shows that the original positive effect of trial becomes almost zero. Still, the interaction **Trial*HighAccuracy** is positive (Model 5: β = 0.006, p < 0.001), reflecting that trust increases over the trial sequence with high accuracy but not with low accuracy (Figure 3). This supports H2 in that trust and reliance decrease less over time (trials) with high model accuracy compared to when it is low.

4.4 Trust and reliance are not affected by different explanation types (H3, H4)

Next, we test whether trust and reliance are higher for human-like AI explanations than for abstract explanations (H3), assuming that human-like explanations "protect" trust and reliance better over time (H4). We first compare explanation types regarding their absolute trust and reliance levels. We find that both trust and reliance are not significantly higher for **HumanExplanation** compared with abstract explanations as our regressions show (table 3). Further, we analyze the effect of explanations over trial, **Trial*HumanExplanation**: we find no significant effect of explanation type over time on reliance (Model 2: interaction effect β = -0.001, p = 0.55) as well or trust (Model 5: interaction effect β = -0.012, p = 0.28). Thus, we reject H3 and H4: human-like explanations do neither affect trust nor reliance differently than abstract explanations. They are furthermore not effectively protecting trust or reliance better over time.

4.5 Human-like explanations elevate reliance for high-accuracy models

As an additional exploration, we test for potential interaction effects of model accuracy and explanation type. Comparing the high-accuracy lines in Figure 3, we find that trust and reliance are higher for human-like

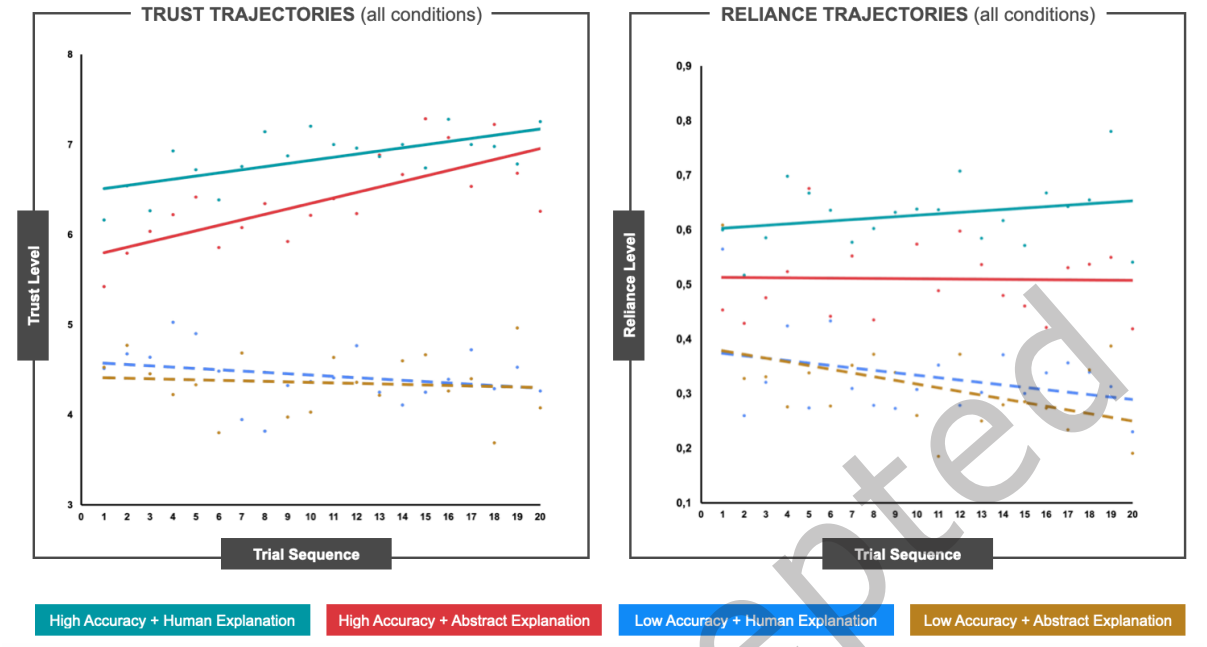

Fig. 3. Development of trust and reliance per condition over the sequence of 20 trial rounds. The dotted lines represent the mean trust and reliance levels, and the solid line represents a fitted line.

explanations. Model 2 shows a significant interaction for reliance for **HumanExplanation*HighAccuracy** (Model 2: β = 0.123, p = 0.03), however, this effect is not seen for trust (Model 5: β = 0.54, p = 0.27). The stronger effect of reliance (in comparison to trust) is also visible in Figure 3. We summarize that reliance is significantly different for highly accurate models with human-like explanations: in contrast to abstract keyword explanations, human-like explanations elevate people's willingness to adapt their advice towards the AI advice.

4.6 Trust and reliance only correlate to a limited extent

We repeatedly measured trust beliefs and reliance behavior for a sequence of 20 legal cases. Running the empty regression models for our two target variables to learn about the proportion of variance, we discover that trust can be attributed to interpersonal differences of participants (ρ = 0.55), whereas for reliance, this level is lower, and we can assume a stronger influence of the study task (ρ = 0.32). A difference in both measurements also shows Figure 3 as trust and reliance trajectories progress differently. Furthermore, we find that trust is significantly different for model accuracy whereas reliance only shows borderline significant effects. A similar discrepancy but in the opposite direction shows the interaction effects of model accuracy and explanation types (H4). In addition, a Pearson's correlation was run to assess the relationship between both measurements. We find a significant moderate correlation between trust and reliance, r(3130) = 0.305, p < 0.001.

4.7 Reliance is influenced by age but not by gender, tech affinity, or legal expertise

In addition to our main predictor variables, we included several covariates in Model 3 (R^2 = 0.14) and Model 6 (R^2 = 0.21). Previous work shows that domain expertise influences trust in AI advice negatively [61] whereas tech

affinity promoted it [74]. We do not find significant effects on **LegalExpertise** for either target variables (Model 3: β = −0.008, p = 0.188; Model 6: β = 0.047, p = 0.35) nor for **TechAffinity** (Model 3: β = −0.011, p = 0.465; Model 6: β = −0.027, p = 0.349). Finally, trust and reliance are neither significantly different for **Female** participants (Model 3: β = - 0.049, p = 0.129; Model 6: β = - 0.123, p = 0.661). However, we find that **Age** significantly affects reliance negatively (Model 3: β = −0.004, p = 0.002), but trust was not affected (Model 6: β = 0.009, p = 0.38). Our finding falls in line with earlier work. For example, Knowles & Hanson [52] found that AI aversion increases with age: as people become older, they grow more resistant or feel to lack sufficient control over technology and, thus, trust technology less. Still, previous research found non-significant covariate effects on trust and reliance [76].

4.8 Prior interactions with AI advice explains subsequent trust and reliance

To explore whether repeated interactions over time impact trust and reliance on AI, we measure whether the interaction in a prior trial round impacts trust and reliance in a subsequent trial round along our 20-trial sequence. We analyze the effect of participants' aggregate prior trust and reliance experience on trust and reliance evaluations of a subsequent trial round with Model 7 (R^2 = 0.25) and Model 10 (R^2 = 0.45), see Table 4. **AggregatePriorReliance** significantly affects reliance for a subsequent trial round (Model 7: β = 0.18, p < 0.001). The same effect is found for **AggregatePriorTrust** (Model 10: β = 0.72, p < 0.001). We summarize that aggregate prior trust and reliance experiences can indeed explain (future) reliance in addition to observing system properties, such as model accuracy.

Second, we test whether participants' immediate prior trust and reliance experience affects trust and reliance in a subsequent trial round with Model 8 (R^2 = 0.21) and Model 11 (R^2 = 0.43), see Table 4. We find that the **ImmediatePriorReliance** behavior significantly affects reliance in a subsequent trial round (Model 8: β = .29, p <0.001). The same effect is found for **ImmediatePriorTrust** beliefs on subsequent trust (Model 11: β = .52, p < 0.001). Comparing the results of immediate prior experiences with the aggregate prior experiences for trust and reliance, we conclude that both can explain trust and reliance. Still, the aggregate prior variables have a slightly stronger effect, more for reliance, β (aggregate) = 0.629, β (immediate) = 0.244, than for trust, β (aggregate) = 0.724, β (immediate) = 0.517.

Finally, we test for potential associations of the immediate prior experience variables of trust and reliance: trust beliefs could be affected by prior reliance behavior and reliance could be affected by prior trust beliefs. Our results are shown in Model 9 (R^2 = 0.23) and Model 12 (R^2 = 0.43), see Table 4. We find that reliance is significantly affected by both **ImmediatePriorReliance** behavior (Model 9: β = 0.26, p < 0.001) as well as by **ImmediatePriorTrust** beliefs (Model 9: β = 0.02, p < 0,001). Furthermore, we find that trust is significantly affected by immediate prior trust (Model 12: β = 0.53, p < 0,001) but immediate prior reliance does not affect trust (Model 12: β = 0.003, p = 0.976).

Summarizing our analysis on the impact of prior trust and reliance experiences, we note that controlling for previous trial interactions strengthens our findings: accounting for time-sensitive variables refines our understanding of the effect on trust and reliance. Not all of the effects of trust and reliance are attributable to model accuracy but to how people observe and learn to follow AI advice. We find that prior interactions over the course of our experiment affect trust and reliance and explain the development of trust and reliance. Those results refine our earlier findings: we uncover that not only does model accuracy impact trust and reliance positively but that people built the latter step-by-step by "reflecting" on prior interactions. As hypothesized, we also find that immediate prior reliance behavior affects trust beliefs significantly, or, in other words, action follows thoughts. However, we do not find the vice-versa effect of prior immediate reliance affecting trust.

Model:	Model 1	Model 2	Model 3
Target Variable:	**Reliance**	**Reliance**	**Reliance**
Predictor Variables:			
HumanExplanation	0.048 (0.030)	− 0.001 (0.049)	− 0.070 (0.490)
HighAccuracy	0.257 (0.030) ***	0.095 (0.052)	0.096 (0.051)
Trial	− 0.001 (0.001)	− 0.004 (0.002) *	− 0.005 (0.002) *
HumanExpl * HighAcc		0.129 (0.060) *	0.123 (0.059) *
Trial * HighAcc		0.008 (0.002) ***	0.008 (0.002) ***
Trial * HumanExpl		− 0.001 (0002)	− 0.001 (0002)
Age			− 0.040 (0.001) **
Female			− 0.049 (0.032)
TechAffinity			− 0.011 (0.015)
LegalExpertise			− 0.008 (0.006)
Model R-Square:	0.12	0.13	0.14

Model:	Model 4	Model 5	Model 6
Target Variable:	**Trust**	**Trust**	**Trust**
Predictor Variables:			
HumanExplanation	0.195 (0.251)	0.060 (0.380)	0.092 (0.383)
HighAccuracy	2.259 (0.248) ***	1.294 (0.402) ***	1.313 (0.406) ***
Trial	0.019 (0.006) **	− 0.003 (0.010)	− 0.003 (0.010)
HumanExpl * HighAcc		0.551 (0.504)	0.541 (0.514)
Trial * HighAcc		0.057 (0.011) ***	0.057 (0.011) ***
Trial * HumanExpl		− 0.012 (0.011)	− 0.012 (0.011)
Age			0.009 (0.011)
Female			− 0.123 (0.280)
TechAffinity			− 0.027 (0.134)
LegalExpertise			0.047 (0.050)
Model R-Square:	0.20	0.21	0.21

$^{*}p$ <0.05, $^{**}p$ <0.01, $^{***}p$ <0.001

Table 3. Nested regression models of Study 1 for the two target variables reliance, measured as weight on advice (Models 1,2,3), and trust, measured with one question item post-trial (Models 4,5,6). The table shows the coefficients (marked with asterisks depending on the significance level). The standard error is given in round brackets.

5 REPLICATION STUDY (STUDY 2)

In addition to our original study (Study 1) discussed in Chapter 4, we further share the results of a conceptual replication study (Study 2), which was set up almost identical in comparison to the original study, with several improvements of the design based on considerations after the initial study. Our aim is to confirm findings from our original study and thus strengthen the credibility of our results.

Model:	Model 7	Model 8	Model 9
Target Variable:	**Reliance**	**Reliance**	**Reliance**
Predictor Variables:			
HumanExplanation	− 0.033 (0.039)	− 0.064 (0.039)	0.066 (0.038)
HighAccuracy	0.180 (0.040) ***	0.150 (0.039) ***	0.123 (0.039) *
HumanExpl * HighAcc	0.090 (0.035) *	0.109 (0.030) ***	0.100 (0.030) *
Trial	− 0.002 (0.002)	− 0.002 (0.002)	− 0.002 (0.002)
Trial * HighAcc	− 0.003 (0.003)	− 0.001 (0.003)	− 0.020 (0.003)
Trial * HumanExpl	0.001 (0.002)	0.003 (0.003)	0.003 (0.030)
AggregatePriorReliance	0.629 (0.050) ***		
AggregatePriorTrust			
ImmediatePriorReliance		0.244 (0.020) ***	0.264 (0.200) ***
ImmediatePriorTrust			0.022 (0.003) ***
Model R-Square:	0.25	0.21	0.23

Model:	Model 10	Model 11	Model 12
Target Variable:	**Trust**	**Trust**	**Trust**
Predictor Variables:			
HumanExplanation	0.080 (0.223)	− 0.041 (0.203)	0.032 (0.210)
HighAccuracy	1.174 (0.230) ***	0.848 (0.207) ***	0.890 (0.214) ***
HumanExpl * HighAcc	0.247 (0.223)	0.253 (0.154)	0.240 (0.161)
Trial	− 0.040 (0.011) **	− 0.004 (0.012)	0.001 (0.013)
Trial * HighAcc	0.001 (0.013)	0.015 (0.014)	0.010 (0.015)
Trial * HumanExpl	− 0.120 (0.012)	0.001 (0.014)	− 0.006 (0.015)
AggregatePriorReliance			
AggregatePriorTrust	0.724 (0.035) ***		
ImmediatePriorReliance			0.003 (0.109)
ImmediatePriorTrust		0.517 (0.016) ***	0.530 (0.017) ***
Model R-Square:	0.45	0.43	0.43

*p <0.05, **p <0.01, ***p <0.001

Table 4. Regression models for the two target variables reliance, measured as weight on advice (Models 7,8,9), and trust, measured as self-reported answers post-trial (Models 10,11,12). The models include the time-sensitive predictor variables "aggregate prior" trust and reliance and "immediate prior" trust and reliance. The table shows the coefficients (marked with asterisks depending on the significance level). The standard error is given in round brackets.

5.1 Study Design & Methodology

5.1.1 Predictor Variables & Study Manipulations. Different than in Study 1, we focused solely on testing model accuracy and its effect on trust and reliance (over time) and, therefore, removed the explanation type manipulation from the study design. However, model accuracy was tested very similarly to study 1 as a two-factorial

manipulation: high model accuracy was defined as a model with estimates that deviated away from the correct jail time (ground truth) maximum +/- 15%, whereas low model accuracy was defined as a model which estimates deviated away from the correct jail time maximum +/- 30%. The reason for testing high versus medium accuracy was to learn whether the effect of accuracy would still remain by comparing models with a more subtle difference in accuracy. In line with Yu et al. [97], we would still expect trust and reliance to increase for high accuracy but not for medium accuracy. Different from the setup in Study 1, all estimates were calculated randomly for all cases and participants within the respective accuracy margins, which limited the potential confounding effects of pre-coded AI estimates. Identical to Study 1, we measured trust and reliance. However, we measured trust more robustly with three question items after each trial (in comparison to one item in the original study): the question items were inspired by a trust scale from Madsen / Madsen & Gregory's [62, 63] *Faith* sub-scale; participants had to indicate to what extent they (1) have faith in the AI advice, (2) believed the AI advice more than themselves, and (3) how confident they were that the AI provided the best solution (7-point Likert scale: 1 = not at all, 7 = completely). A scale analysis showed excellent levels of internal consistency (α = .95). Again different from the original study, we measured trust once more after completing all legal cases with a post-trial trust scale consisting of three question items inspired by the *Perceived Reliability* and *Perceived Technical Competence* sub-scales, again from Madsen / Madsen & Gregor [62, 63]. We asked participants to what extent they agreed that (1) the AI seemed trustworthy to them, (2) the AI performance was consistent over all trials, and (3) the AI advice was just as good as that of a highly competent human (7-point Likert scale: 1 = fully disagree, 7 = fully agree). A scale analysis showed good levels of internal consistency (α = 0.82). Study 1 and 2 included the same covariates: age, gender, legal expertise, and technological affinity. In addition to these, we included the covariates religiousness, conscientiousness, and socio-economic question items in our post-experimental questionnaire. We did not include any of these covariates in the final analysis as they showed non-significant effects in the initial inspection (similar to our findings in the original study).

5.1.2 Procedure. Study 1 and 2 were identical in the procedure of the study task except for two minor adaptations. First, we omitted the interactive visual calculating element (waiting cursor) between participants' initial estimate and seeing the AI estimate (see Figure 2). Furthermore, we resolved each legal case in more detail (Step 5): participants learned about the correct jail time and were reminded of their second (final) estimate and the AI's estimate. As the two studies were based on identical tasks, we again retrieved 20 criminal law cases from the Dutch database *De Rechtspraak* [23]. However, we chose 20 new cases for the replication study. In comparison to Study 1, we selected cases that, in total, offered a rather homogeneous jail time between 4 and 30 months (whereas the original study selected cases with jail time between 4 and 120 months).

5.1.3 Participants. For our analysis of Study 2, we approximate the sample size with an a priori power analysis (ANOVA fixed effects, main effects, and interactions; effect size f = 0.25, α error = 0.05, power = 0.9, two groups) based on our manipulation factor model accuracy (two levels), which resulted in N=171 required participants. The sample size can be interpreted as an approximation of the necessary sample size under the assumption that there is a low correlation between the repeated measures. We base our calculation on similar studies [75, 76, 90, 97] with repeated trials with medium effect sizes (0.25) and comparable sample sizes. As the study was intended to test additional assumptions outside of this paper's scope, we extended the sample size and recruited N=211 participants via Prolific [78]. In addition, we recruited N=10 participants for a pilot study. We selected participants by age (>18 years), residency (UK), and Prolific approval rate (>95%) and excluded participants who participated in similar experiments (among others, participants from the original study). The study was conducted in May 2023. Participants were compensated with £12.52 per hour. They finished the experiment in 21 minutes (median time).

5.1.4 Statistical Analysis. One reason to run a replication study was to support the validity of our results from original STUDY 1. As the design of both studies is almost identical (for more details, see Chapter 5.1), we also approach the analysis of our replication study the same way. The main difference is that the replication study does not include the explanation type manipulation and disregards the analysis of most of the covariates done in STUDY 1, for example, gender, age, legal expertise, tech affinity[5]. To summarize, we discuss the effect of model accuracy on trust and reliance based on several multi-level regression models (repeated trials nested within participants). The regression models of the replication study are labeled with the letter R as the suffix, for example, *Model 7R*. All models can be found in Table 5.

5.2 Results

From N=211 participants (including pilot study participants), we analyzed data from N=200 participants. We excluded one participant due to incomplete data and the 10 participants from the pilot testing. On average, participants were 40.1 years old (Min: 18, Max: 79, *SD* = 13.43), 61% identified as female (37% as male, 1% as non-binary, and 1% preferred not to indicate any gender). On average, participants indicated a somewhat lower level of legal expertise of *M*= 3.67 (Min: 1, Max: 10, *SD* = 2.35) than in STUDY 1 and a medium level of tech affinity of *M*= 2.63 (Min: 1, Max: 5, *SD* = 0.97). Preparing our data for the analysis of our predictor variable reliance, we find that from a total of 4,000 observations (200 participants multiplied by 20 legal case trials), WoA was 1 for 499 of those observations (= participants adopted the AI estimate for their final estimate); WoA was 0 for 965 observations (= participants initial estimate was identical to their final estimate); WoA was missing for 154 cases, which happened for WoA calculations divided by 0 (= participants initial estimate and AI estimate was identical). For our final data set, we defined WoA margins from -2 to +2, which resulted in excluding 13 outlier observations.

Before discussing the results from our regression models, we compare mean levels of reliance (across trials), trust (across trials), and post-trial trust by (accuracy) groups with a two-sample t-test with equal variances. We find that reliance is significantly higher for high accuracy (M = 0.69, SD = 0.40) than for medium accuracy (M = 0.48, SD = 0.42), t (3844) = −15.87, p < 0.001. In line, we find that trust is significantly higher for high accuracy (M = 5.46, SD = 1.49) than for medium accuracy (M = 4.02, SD = 1.53), t (3998) = -30.1, p < 0.001. Similarly, post-trial trust is higher for high accuracy (M = 4.09, SD = 0.78) than for medium accuracy (M = 3.16, SD = 0.77), t (3998) = -38.20, p < 0.001.

In line with findings from our original STUDY 1, we find that trust and reliance are higher for **HighAccuracy** (Model 7R: β = 0.090, p < 0.001; Model 10R: β = 0.436, p < 0.001), see Table 5). In addition, reliance grows positively significantly over time for high accuracy, **Trial*HighAccuracy** (Model 7R: β = 0.006, p = 0.001) but not trust (Model 10R: β = 0.003, p = 0.482), which therefore partly confirms findings from STUDY 1.

In line with procedures for our original study, we included predictor variables to test whether prior trust and reliance experiences along the 20-trial sequence affected reliance or trust in subsequent trials. We distinguish between aggregate prior experiences that account for all past trial rounds of a participant and the trial round that was happening immediately before a subsequent one. We again find that participants are significantly affected by their **AggregatePriorReliance** behavior (Model 7R: β = 0.353, p > 0.001). We find the same for trust beliefs: participants are significantly affected by their **AggregatePriorTrust** beliefs (Model 10R: β = 0.860, p < 0.001). We find the same for **ImmediatePriorReliance** behavior as it significantly affects participant's reliance (Model 8R: β =0.200, p <0.001). In addition, **ImmediatePriorTrust** beliefs significantly affect participants' trust (Model 11R: β = 0.818, p > 0.001).

Finally, we tested whether immediate prior reliance affected trust in a subsequent trial and vice versa (Models 9R and 12R). We find that reliance is significantly affected by **ImmediatePriorReliance** (Model 9R: β = 0.149, p

[5]We did not include the covariates age, gender, tech affinity or legal expertise in our replication study models as they did not produce a significant outcome in previous analysis steps.

< 0.001) as well as by **ImmediatePriorTrust** (Model 9R: β = 0.061, p < 0.001); this is in line with findings in our original study. Furthermore, we find that trust is significantly affected by **ImmediatePriorTrust** (Model 12R: β = 0.810, p < 0.001) and trust is also significantly affected by **ImmediatePriorReliance** (Model 12R: β = 0.106, p = 0.001). This slightly differs from the original study, where only immediate prior trust affected trust, but not reliance.

In conclusion, our results of Study 2 predominantly replicate the findings of the original Study 1: trust and reliance are higher for high-accuracy models, and reliance increases over time (= over the trial sequence) with high accuracy. With additional analyses, we find that prior trust and reliance experiences explain reliance behavior and trust beliefs in subsequent trials. That is, if participants trusted or relied on AI advice before, it is likely that they will continue relying on and expressing trust in the AI output. Interestingly, we find that trust and reliance are also interconnected regarding these temporal effects: trust in a prior trial round influences reliance behavior in a subsequent trial, and reliance in a prior trial round also influences trust beliefs in a subsequent trial round.

6 DISCUSSION

Our two studies provide new insights into how trust develops over a repeated decision-making task with AI support. Our results confirm prior findings of the HAI literature but add further details about the interplay between our two experimental conditions, model accuracy, explanation type, and the correlation of trust beliefs and trust behavior (reliance). Furthermore, they suggest that prior trust and reliance experiences over a sequential task scenario impact subsequent trust and reliance and, therefore, the development of trust in AI.

Our results from both studies confirm our hypothesis that trust and reliance are higher for more accurate AI models (H1). Even if people's ability to distinguish a good from a bad one may be taken for granted or at least desired, earlier work did not always find appropriate trust in AI (advice). Moreover, we found self-reported trust increasing over time for highly accurate AI advice as proposed in H2: participants picked up the competency of the AI system. They followed its advice more as time progressed. Results for reliance were less strong: we found reliance stabilizing over time without increasing too much. We could partly confirm those findings with our replication study as reliance also significantly increased over the trial sequence in the high accuracy condition. One interpretation of why trust was not found to be significantly increasing over time could be explained by the high accuracy of the AI advice: participants may have been satisfied with the high accuracy of the AI from an early stage of the experiment, which may have led to a ceiling effect in the high accuracy condition. Comparing our studies with previous work, we find our results match those of Yu et al. [97] and their experimental conditions of 80% and 90% accuracy, where trust also increased over time (but not for 70% accuracy). In addition, Papenmeier et al. [75] measured trust in AI for three different accuracy levels (high, medium, low) with results showing that accuracy (vs. explanations) impacted trust overall the most. Their latest work [76] confirms that high accuracy increases trust and that participants can pick up on a sufficiently competent algorithmic system. Based on current knowledge, we propose that people can tell a good from a bad model when model accuracy is high.

Results of Study 1 reveal no significant effect of the type of explanation on trust and reliance: participants were not willing to rely more on AI advice that was supported by a human-like explanation than AI advice supported by an abstract one. Results were non-significant for absolute trust and reliance values and for the development over the trial sequence. We hypothesized that human-like explanations would portray the advice of an AI model as more sophisticated and, thus, would be perceived as more trustworthy. Our argumentation followed most studies in the HAI literature claiming that explanations increase trust and reliance in computer models, cf. [58]. For example, we assume that the design of explanations was not suitable for increasing trust and reliance, for example, as they added no clear benefit other than repeating or highlighting words from the legal case text. The presentation of the explanations may also have lacked ecological validity, as we did not use real explanations using XAI tools such as LIME. Other (X)AI properties, such as model confidence or visualizations

Model:	Model 7R	Model 8R	Model 9R
Target Variable:	**Reliance**	**Reliance**	**Reliance**
Predictor Variables:			
HighAccuracy	0.090 (0.031) ***	0.110 (0.030) ***	0.060 (0.030)
Trial	− 0.001 (0.002)	− 0.001 (0.002)	− 0.001 (0.002)
Trial * HighAcc	0.006 (0.002) **	0.007 (0.002) **	0.004 (0.002)
AggregatePriorReliance	0.353 (0.031) ***		
AggregatePriorTrust			
ImmediatePriorReliance		0.200 (0.015) ***	0.149 (0.015) ***
ImmediatePriorTrust			0.061 (0.004) ***
Model R-Square:	0.16	0.11	0.16

Model:	Model 10R	Model 11R	Model 12R
Target Variable:	**Trust**	**Trust**	**Trust**
Predictor Variables:			
HighAccuracy	0.436 (0.072) ***	0.311 (0.064) ***	0.300 (0.065) ***
Trial	0.003 (0.003)	− 0.001 (0.004)	− 0.001 (0.004)
Trial * HighAcc	0.003 (0.005)	0.001 (0.005)	0.001 (0.005)
AggregatePriorReliance			
AggregatePriorTrust	0.860 (0.018) ***		
ImmediatePriorReliance			0.106 (0.033) **
ImmediatePriorTrust		0.818 (0.009) ***	0.810 (0.010) ***
Model R-Square:	0.76	0.73	0.74

$^{*}p$ <0.05, $^{**}p$ <0.01, $^{***}p$ <0.001

Table 5. Nested regression models for the two target variables reliance, measured as weight on advice (models 7R,8R,9R), and trust, measured with question items post-trial (Models 10R,11R,12R). The models include the time-sensitive predictor variables "aggregate prior" trust and reliance and "immediate prior" trust and reliance. The table shows the coefficients (marked with asterisks depending on the significance level). The standard error is given in round brackets. The model named includes "R" to differentiate the models from the replication of study 2 and the models from the original study 1.

[48], could have resulted in different outcomes. Papenmeier et al. [75] concluded that not giving an explanation resulted in better or equal trust compared to giving an explanation. They reason that their approach (highlighting words in the text) may not have added any further trustworthiness to enhance understanding and, thus, increase willingness to follow AI advice.

Although our explanation did not affect trust and reliance as such, we found that reliance (but not trust) was boosted for an accurate model with human-like explanations. We interpret that the sophisticated explanations corroborate high model accuracy and enable participants to understand its processes better and follow AI advice. Interestingly, this counter-argues the findings of Papenmeier et al. [76]: trust was not significantly higher in their most accurate condition (faithful explanations in a high-accuracy model). They argue that "faithful explanations

do not necessarily imply meaningfulness in the eyes of the user" (p. 26) as these are based on statistical relations rather than causal information. Our results are consistent with the fact that participants found the human-like explanations sufficiently supportive, which could be because the explanations were similarly constructed in comparison to their own evaluation strategies (selecting key issues, weighing and summarizing arguments for a final assessment). Overall, we welcome the outcome that unfaithful explanations (seemingly logical, human-like explanations paired with low-accuracy AI advice) did not trick participants into sub-optimal estimates; this is also in line with the results of Knijnenburg and Willemsen [51].

Measuring trust and reliance after each trial, we found that self-reported trust significantly increased over the sequence of 20 trials whereas reliance stabilized. One explanation could lie in the different trust dimensions: trust beliefs could be described as rather (stable) person-consistent, whereas trust behavior (reliance) would be more trial-dependent and less stable (a basis for this assumption gives our comparison of variance of trust and reliance). A practical reason could lie in the setup of our trials: whereas reliance was defined during the participants' estimation, trust was assessed after learning how the AI overall did. Although we purposely intended to let participants know how the AI was doing after each trial, this could be one of the reasons why trust was expressed more clearly than reliance. Although we could confirm our results with our replication study, the question of an appropriate measurement remains: are people able to express trust adequately, or is relying on AI advice a stronger predictor, c.f. [7, 81]?

Finally, we explored the impact of prior interactions with AI advice and their impact on trust and reliance in subsequent trials, as was analyzed within our 20-trial sequence. We assumed that not only model accuracy and explanation type may affect trust and reliance but that the latter could be explained by "learning" over time. For example, the research community overly agrees that first impressions of an AI model can set the tone for subsequent behavior or beliefs [25, 86]. Thus, when researching repeated interactions, it is important to acknowledge that people do not act isolated but change and adapt over time and consider gained knowledge, which is still often neglected in most statistical analyses, even when researchers have the repeated data available. We found that both trust and reliance were significantly affected by prior interactions: subsequent trials were influenced by both the immediate and the aggregate prior trust and reliance interactions: participants observed how the AI did in previous trial rounds (and how they did themselves with support of it with estimating jail times), which built up trust and influenced their subsequent trust and reliance evaluations. In line with this, we found that participants' previously built trust beliefs had an effect on their willingness to rely on AI advice. However, prior reliance did not significantly affect trust beliefs in a subsequent trial (in Study 1). The latter findings somewhat support our notion that trust is more stable over time whereas reliance is based on the specific case at hand. Although we also find effects in both directions (prior reliance affecting subsequent trust) in our replication study, we can conclude that the effects of prior trust on subsequent reliance are bigger than vice versa based on our found effect sizes.

We have evidence to support the claim that people rely on and trust more when model accuracy is high **(H1)**. Furthermore, reliance stabilizes, and trust increases over time with high accuracy **(H2)**. Explanation type, however, does not affect trust and reliance **(H3, H4)**: both measures were not affected, neither as absolute value nor over time. Some additional analysis showed that human-like explanations boost reliance on high-accuracy models. Therefore, we can answer our research questions: trust and reliance are higher with high model accuracy, and trust also positively develops over time as proposed in **RQ1**. Regarding **RQ2**, we did not find evidence that human-like explanations affect trust and reliance differently than abstract ones. Finally, we explored the impact of prior trial interactions regarding trust and reliance on subsequent trust and reliance effects **(RQ3)** and found that prior trust and reliance interactions significantly influence trust beliefs and behavior in subsequent trials. This seems logical as interactions with AI systems do not happen in a vacuum, but people naturally learn to interact with them over time. We stipulate that evaluating participants' prior interactions with AI advice provided us with a more detailed picture of how trust and reliance develop over time. Accordingly, we will continue measuring

these effects and hope our study will inspire other researchers to do the same. Understanding how trust and reliance develop may be an important aspect of reaching appropriate trust levels in various HAI applications in the future.

7 LIMITATIONS AND FUTURE WORK

Our work uncovered insights into trust and reliance development with a legal decision-making HAI scenario, shedding light on the effects of model accuracy and explanation type and, furthermore, accounting for prior trust and reliance experiences on subsequent trust and reliance evaluations while receiving AI advice. Despite the significant effects of both presented studies, we would like to discuss some aspects that may limit the results of our study.

Although we tried to test trust with a more realistic experimental design (legal decision-making with real data), the question remains whether we were more successful in that compared to prior work. We could argue that the procedure, that is, resolving each case at the end of each trial round, does not depict a real-life procedure. On the other hand, we could claim that it nevertheless teaches inexperienced participants to understand the AI model better this way. In addition, real decision-making involves a more comprehensive, iterative approach - a shortcoming that some study participants with a legal background commented on. Furthermore, there may have been additional challenges with our legal task. First, jail sentencing standards are country-specific, and so could participants' perceptions, depending on their cultural coinage, be different overall. Second, legal rulings are not neutral by themselves, as individual legal experts decide upon them; therefore, we cannot assume that the jail times of our cases represent an entirely objective ground truth. Thus, participants had to estimate in a noisy environment. One solution to counteract some bias would have been to define jail times based on the mean (or median) of similar cases in the database. Another overall point of critique could be those study participants did not believe in involving AI in (moral) legal decision-making in the first place, which could have affected their engagement and, therefore, our results. We acknowledge the importance of discussing various ethical issues regarding the use of AI in a legal context but want to emphasize that they exceed the scope of this work, which was to observe trust beliefs and behavior in response to receiving AI advice.

To ensure an adequate understanding of how people trust in and rely on AI advice, we measured two trust dimensions: trust beliefs ("trust") and trust behavior ("reliance"). We found that trust and reliance slightly differed in our 20-trial study. We also found that trust and reliance correlate, however, both dimensions also differ in the extent to which trust beliefs can predict subsequent reliance behavior and vice versa. Although our replication study provided results on trust and reliance that were consistent with those of our original study, we remain cautious about the validity and reliability of trust measures and the corresponding interpretation of our study results. We measured reliance based on the adoption of AI advice from study participants using the concept of Weight-on-Advice, an established and widely used method to measure people's trust behavior. However, we measured trust based on participants' self-reports using only short (one-item / three-item) questionnaires. Future studies might benefit from either using more elaborate scales or interviewing participants post-experiment to understand their considerations during the study better.

Although we claim to learn more about the development of trust with our study, it is unclear whether 20 successive interactions in a single session of 20 minutes is sufficient. Other work [90] offers a compromise and measured trust development with separate follow-up sessions. Thus, we intend to measure human-AI collaboration over an extended period of time in the future to draw more robust conclusions about the dynamics and effects of trust over time.

Finally, we want to note that we did not employ an operating AI model for our study but simulated the AI interaction based on pre-determined estimates. This allowed us to remain in control over our study treatments without dealing with potentially unpredictable AI behavior. As it was our focus to study human behavior

concerning AI advice, we believe that this could also be observed in an artificial surrounding and, furthermore, allow us to replicate our experiment more easily, as we did with STUDY 2. We acknowledge the fact that simulations cannot simply be generalized as we were not able to capture the full complexity of a decision-making scenario. Thus, the ecological validity of our results may be limited. Nevertheless, we hope our work inspires practitioners to build AI systems based on a better understanding of human (trust) behavior.

8 CONCLUSION

Without a doubt, there are now numerous domains that could benefit from the use of AI. Accordingly, there is a growing desire to combine human talent and AI capabilities. However, uncertainties about how AI-supported human decision-making needs to be designed still exist. With our study, we were able to demonstrate that model accuracy has significant effects on people's trust and reliance development. We found no differences in the way AI advice was explained. We found that human-like explanations could elevate trust when model performance was high. Furthermore, we controlled for the effect of prior trust and reliance of participants to see whether prior interactions would affect future beliefs or behavior. We indeed found that prior experiences positively affect trust and reliance on subsequent trials. We also see that prior trust affects trust and reliance in subsequent trial rounds. With our results, we were able to contribute new insights related to the development of trust and reliance in the context of a real-life task, especially by considering time-sensitive variables. We want to continue exploring how system performance, people's expectations, and decision-making circumstances influence trust. Our ultimate goal is to enable people to successfully calibrate their trust when systems act outside of the expected frame, similar to how human counterparts sometimes do.

ACKNOWLEDGMENTS

We thank Luc Siecker, Jane Deijnen, Milo Simons, Lorea Ros, and Ruben van der Werf for their help in conducting the original study. Furthermore, we thank Sanderijn van Loosdrecht, Madalina Rogozan, Marit van der Lit, and Yifei Mao for their help in conducting the replication study. Finally, we would like to express our gratitude to the European Supply Chain Forum (ESCF), the Department of Industrial Engineering and Innovation Sciences (IE&IS), the Eindhoven Artificial Intelligence Systems Institute (EAISI), and the Logistics Community Brabant for sponsoring the research project *AI Planner of the Future*, which is funding the PhD project "Trust in AI over time" and thus supporting this and future work.

REFERENCES

[1] Naomi Aoki. 2021. The importance of the assurance that "humans are still in the decision loop" for public trust in artificial intelligence: Evidence from an online experiment. *Computers in Human Behavior* 114 (2021), 106572.

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[3] Phoebe E Bailey, Tarren Leon, Natalie C Ebner, Ahmed A Moustafa, and Gabrielle Weidemann. 2023. A meta-analysis of the weight of advice in decision-making. *Current Psychology* 42, 28 (2023), 24516–24541.

[4] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. 2021. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11405–11414.

[5] Rachel Baumsteiger and Jason T Siegel. 2019. Measuring prosociality: The development of a prosocial behavioral intentions scale. *Journal of personality assessment* 101, 3 (2019), 305–314.

[6] Sarah Bayer, Henner Gimpel, and Moritz Markgraf. 2022. The role of domain expertise in trusting and following explainable AI decision support systems. *Journal of Decision Systems* 32, 1 (2022), 110–138.

[7] Michaela Benk, Suzanne Tolmeijer, Florian von Wangenheim, and Andrea Ferrario. 2022. The Value of Measuring Trust in AI-A Socio-Technical System Perspective. *arXiv preprint arXiv:2204.13480* (2022).

[8] Benedikt Berger, Martin Adam, Alexander Rühr, and Alexander Benlian. 2021. Watch me improve—algorithm aversion and demonstrating the ability to learn. *Business & Information Systems Engineering* 63, 1 (2021), 55–68.

[9] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.
[10] Christopher Burr, Nello Cristianini, and James Ladyman. 2018. An analysis of the interaction between intelligent software agents and human users. *Minds and machines* 28, 4 (2018), 735–774.
[11] Francesca Cabiddu, Ludovica Moi, Gerardo Patriotta, and David G Allen. 2022. Why do users trust algorithms? A review and conceptualization of initial trust and trust over time. *European management journal* 40, 5 (2022), 685–706.
[12] Christopher S Calhoun, Philip Bobko, Jennie J Gallimore, and Joseph B Lyons. 2019. Linking precursors of interpersonal trust to human-automation trust: An expanded typology and exploratory experiment. *Journal of Trust Research* 9, 1 (2019), 28–46.
[13] Noah Castelo and Adrian F Ward. 2021. Conservatism predicts aversion to consequential Artificial Intelligence. *Plos one* 16, 12 (2021), e0261467.
[14] Alvaro Chacon, Edgar E Kausel, and Tomas Reyes. 2022. A longitudinal approach for understanding algorithm use. *Journal of Behavioral Decision Making* (2022).
[15] Chih-Yang Chao, Tsai-Chu Chang, Hui-Chun Wu, Yong-Shun Lin, and Po-Chen Chen. 2016. The interrelationship between intelligent agents' characteristics and users' intention in a search engine by making beliefs and perceived risks mediators. *Computers in Human Behavior* 64 (2016), 117–125.
[16] Jessie YC Chen, Michael J Barnes, Anthony R Selkowitz, Kimberly Stowers, Shan G Lakhmani, and Nicholas Kasdaglis. 2016. Human-autonomy teaming and agent transparency. In *Companion Publication of the 21st International Conference on Intelligent User Interfaces*. 28–31.
[17] Manolis Chiou, Faye McCabe, Markella Grigoriou, and Rustam Stolkin. 2021. Trust, shared understanding and locus of control in mixed-initiative robotic systems. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 684–691.
[18] Leah Chong, Guanglu Zhang, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan. 2022. Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior* 127 (2022), 107018.
[19] Angèle Christin. 2017. Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society* 4, 2 (2017), 2053951717718855.
[20] Eric Chu, Deb Roy, and Jacob Andreas. 2020. Are visual explanations useful? a case study in model-in-the-loop prediction. *arXiv preprint arXiv:2007.12248* (2020).
[21] Cristina Conati, Oswald Barral, Vanessa Putnam, and Lea Rieger. 2021. Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial intelligence* 298 (2021), 103503.
[22] Karl de Fine Licht and Bengt Brülde. 2021. On defining "Reliance" and "Trust": Purposes, conditions of adequacy, and new definitions. *Philosophia* 49 (2021), 1981–2001.
[23] de Rechtspaak. 2022. *de Rechtspraak Website*. https://www.rechtspraak.nl/
[24] Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister. 2019. Hybrid intelligence. *Business & Information Systems Engineering* 61, 5 (2019), 637–643.
[25] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. 2013. Impact of robot failures and feedback on real-time trust. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 251–258.
[26] Berkeley J Dietvorst and Soaham Bharti. 2020. People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological science* 31, 10 (2020), 1302–1314.
[27] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.
[28] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management science* 64, 3 (2018), 1155–1170.
[29] Mary Dzindolet, Linda Pierce, Scott Peterson, Lori Purcell, Hall Beck, and Hall Beck. 2002. The influence of feedback on automation use, misuse, and disuse. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 46. SAGE Publications Sage CA: Los Angeles, CA, 551–555.
[30] Connor Esterwood and Lionel P Robert. 2021. Do you still trust me? human-robot trust repair strategies. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 183–188.
[31] Md Abdullah Al Fahim, Mohammad Maifi Hasan Khan, Theodore Jensen, Yusuf Albayram, and Emil Coman. 2021. Do integral emotions affect trust? The mediating effect of emotions on trust in the context of human-agent interaction. In *Designing Interactive Systems Conference 2021*. 1492–1503.
[32] Rino Falcone and Cristiano Castelfranchi. 2004. Trust dynamics: How trust is influenced by direct experiences and by trust itself. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004.* IEEE, 740–747.
[33] Xiaocong Fan, Sooyoung Oh, Michael McNeese, John Yen, Haydee Cuevas, Laura Strater, and Mica R Endsley. 2008. The influence of agent reliability on trust in human-agent collaboration. In *Proceedings of the 15th European conference on Cognitive ergonomics: the*

ergonomics of cool interaction. 1–8.
[34] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 41, 4 (2009), 1149–1160.
[35] Juliana Jansen Ferreira and Mateus Monteiro. 2021. The human-AI relationship in decision-making: AI explanation to support people on justifying their decisions. *arXiv preprint arXiv:2102.05460* (2021).
[36] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine* 4, 1 (2021), 1–8.
[37] Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* 14, 2 (2020), 627–660.
[38] Sanford C Goldberg. 2020. Trust and reliance. *The routledge handbook of trust and philosophy* (2020), 97–108.
[39] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual visual explanations. In *International Conference on Machine Learning*. PMLR, 2376–2384.
[40] William M Grove and Paul E Meehl. 1996. Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, public policy, and law* 2, 2 (1996), 293.
[41] Feyza Merve Hafizoğlu and Sandip Sen. 2019. Understanding the influences of past experience on trust in human-agent teamwork. *ACM Transactions on Internet Technology (TOIT)* 19, 4 (2019), 1–22.
[42] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.
[43] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. 2016. User trust in intelligent systems: A journey over time. In *Proceedings of the 21st international conference on intelligent user interfaces*. 164–168.
[44] Michael C Horowitz, Lauren Kahn, Julia Macdonald, and Jacquelyn Schneider. 2023. Adopting AI: how familiarity breeds both trust and contempt. *AI & society* (2023), 1–15.
[45] Antoine Hudon, Théophile Demazure, Alexander Karran, Pierre-Majorique Léger, and Sylvain Sénécal. 2021. Explainable artificial intelligence (XAI): how the visualization of AI predictions affects user cognitive load and confidence. In *Information Systems and Neuroscience: NeuroIS Retreat 2021*. Springer, 237–246.
[46] Patricia K Kahr, Gerrit Rooks, Martijn C Willemsen, and Chris CP Snijders. 2023. It seems smart, but it acts stupid: Development of trust in ai advice in a repeated legal decision-making task. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 528–539.
[47] Uday Kamath and John Liu. 2021. *Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning*. Springer.
[48] Alexander John Karran, Théophile Demazure, Antoine Hudon, Sylvain Senecal, and Pierre-Majorique Léger. 2022. Designing for Confidence: The Impact of Visualizing Artificial Intelligence Decisions. *Frontiers in Neuroscience* 16 (2022).
[49] Rabia Fatima Khan and Alistair Sutcliffe. 2014. Attractive agents are more persuasive. *International Journal of Human-Computer Interaction* 30, 2 (2014), 142–150.
[50] Taenyun Kim and Hayeon Song. 2021. How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics* 61 (2021), 101595.
[51] Bart Knijnenburg and Martijn Willemsen. 2016. Inferring Capabilities of Intelligent Agents from Their External Traits. *ACM Transactions on Interactive Intelligent Systems* 6 (11 2016), 1–25. https://doi.org/10.1145/2963106
[52] Bran Knowles and Vicki L. Hanson. 2018. The Wisdom of Older Technology (Non)Users. *Commun. ACM* 61, 3 (feb 2018), 72–77. https://doi.org/10.1145/3179995
[53] Spencer C Kohn, Daniel Quinn, Richard Pak, Ewart J De Visser, and Tyler H Shaw. 2018. Trust repair strategies with self-driving vehicles: An exploratory study. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 62. Sage Publications Sage CA: Los Angeles, CA, 1108–1112.
[54] Moritz Körber, Eva Baseler, and Klaus Bengler. 2018. Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied ergonomics* 66 (2018), 18–31.
[55] Maier Fenster1and Inon Zuckerman2and Sarit Kraus. 2012. Guiding user choice during discussion by silence, examples and justifications. In *ECAI 2012: 20th European Conference on Artificial Intelligence*, Vol. 242. IOS Press, 330.
[56] John Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (1992), 1243–1270.
[57] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.
[58] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2119–2128.
[59] Brian Y Lim, Qian Yang, Ashraf M Abdul, and Danding Wang. 2019. Why these explanations? Selecting intelligibility types for explanation goals.. In *IUI Workshops*.

[60] Tyler J Loftus, Patrick J Tighe, Amanda C Filiberto, Philip A Efron, Scott C Brakenridge, Alicia M Mohr, Parisa Rashidi, Gilbert R Upchurch, and Azra Bihorac. 2020. Artificial intelligence and surgical decision-making. *JAMA surgery* 155, 2 (2020), 148–158.
[61] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.
[62] Maria Madsen. 2000. *The Development of a Psychometric Instrument for human-computer trust.*
[63] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *11th australasian conference on information systems*, Vol. 53. Citeseer, 6–8.
[64] JB Manchon, Mercedes Bueno, and Jordan Navarro. 2021. Calibration of Trust in Automated Driving: A Matter of Initial Level of Trust and Automated Driving Style? *Human Factors* (2021), 00187208211052804.
[65] Dietrich Manzey, Juliane Reichenbach, and Linda Onnasch. 2012. Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making* 6, 1 (2012), 57–87.
[66] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.
[67] D Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and validating trust measures for e-commerce: An integrative typology. *Information systems research* 13, 3 (2002), 334–359.
[68] D Harrison McKnight, Larry L Cummings, and Norman L Chervany. 1998. Initial trust formation in new organizational relationships. *Academy of Management review* 23, 3 (1998), 473–490.
[69] Marieke Möhlmann and Lior Zalmanson. 2017. Hands on the wheel: Navigating algorithmic management and Uber drivers'. In *Autonomy', in proceedings of the international conference on information systems (ICIS), Seoul South Korea*. 10–13.
[70] Ilja Nastjuk, Bernd Herrenkind, Mauricio Marrone, Alfred Benedikt Brendel, and Lutz M Kolbe. 2020. What drives the acceptance of autonomous driving? An investigation of acceptance factors from an end-user's perspective. *Technological Forecasting and Social Change* 161 (2020), 120319.
[71] Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D Ragan. 2019. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 97–105.
[72] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 112–121.
[73] Kazuo Okamura and Seiji Yamada. 2020. Adaptive trust calibration for human-AI collaboration. *Plos one* 15, 2 (2020), e0229132.
[74] Atte Oksanen, Nina Savela, Rita Latikka, and Aki Koivula. 2020. Trust toward robots and artificial intelligence: An experimental approach to human–technology interactions online. *Frontiers in Psychology* 11 (2020), 568256.
[75] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652* (2019).
[76] Andrea Papenmeier, Dagmar Kern, Gwenn Englebienne, and Christin Seifert. 2022. It's Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI. *ACM Transactions on Computer-Human Interaction (TOCHI)* 29, 4 (2022), 1–33.
[77] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.
[78] Prolific.co. 2022. *Prolific Research Platform.* https://www.prolific.co/
[79] Timothy M Rawson, Raheelah Ahmad, Christofer Toumazou, Pantelis Georgiou, and Alison H Holmes. 2019. Artificial intelligence can improve decision-making in infection management. *Nature Human Behaviour* 3, 6 (2019), 543–545.
[80] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
[81] Nicolas Scharowski, Sebastian AC Perrig, Nick von Felten, and Florian Brühlmann. 2022. Trust and Reliance in XAI–Distinguishing Between Attitudinal and Behavioral Measures. *arXiv preprint arXiv:2203.12318* (2022).
[82] F David Schoorman, Roger C Mayer, and James H Davis. 2007. An integrative model of organizational trust: Past, present, and future. , 344–354 pages.
[83] Navya Nishith Sharan and Daniela Maria Romano. 2020. The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon* 6, 8 (2020), e04572.
[84] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies* 146 (2021), 102551.
[85] Donghee Shin, Bu Zhong, and Frank A Biocca. 2020. Beyond user experience: What constitutes algorithmic experiences? *International Journal of Information Management* 52 (2020), 102061.
[86] Keng Siau and Weiyu Wang. 2018. Building trust in artificial intelligence, machine learning, and robotics. *Cutter business technology journal* 31, 2 (2018), 47–53.
[87] Janet A. Sniezek and Lyn M. Van Swol. 2001. Trust, Confidence, and Expertise in a Judge-Advisor System. *Organizational Behavior and Human Decision Processes* 84, 2 (2001), 288–307. https://doi.org/10.1006/obhd.2000.2926

[88] Siddharth Swaroop, Zana Buçinca, and Finale Doshi-Velez. 2023. Adaptive interventions for both accuracy and time in AI-assisted human decision making. *arXiv preprint arXiv:2306.07458* (2023).
[89] Andrea Tocchetti and Marco Brambilla. 2022. The Role of Human Knowledge in Explainable AI. *Data* 7, 7 (2022), 93.
[90] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. Second chance for a first impression? Trust development in intelligent system interaction. In *Proceedings of the 29th ACM Conference on user modeling, adaptation and personalization*. 77–87.
[91] Ning Wang, David V Pynadath, and Susan G Hill. 2016. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 109–116.
[92] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th international conference on intelligent user interfaces*. 318–328.
[93] Adrian Weller. 2019. Transparency: motivations and challenges. In *Explainable AI: interpreting, explaining and visualizing deep learning*. Springer, 23–40.
[94] Daniel Wessel, Christiane Attig, and Thomas Franke. 2019. ATI-S-an Ultra-Short scale for assessing affinity for technology interaction in user studies. In *Proceedings of Mensch und Computer 2019*. 147–154.
[95] X Jessie Yang, Christopher Schemanske, and Christine Searle. 2021. Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation. *arXiv preprint arXiv:2107.07374* (2021).
[96] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.
[97] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd international conference on intelligent user interfaces*. 307–317.
[98] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 295–305.